

CompAdaGrad: A Compressed, Complementary, Computationally-Efficient Adaptive Gradient Method

Nishant A. Mehta*
Centrum Wiskunde & Informatica
mehta@cw.nl

Alistair Rendell
Australian National University
Alistair.Rendell@anu.edu.au

Anish Varghese
Australian National University
anish.varghese@anu.edu.au

Christfried Webers†
Data61 & Australian National University
christfried.webers@data61.csiro.au

Abstract

The adaptive gradient online learning method known as AdaGrad has seen widespread use in the machine learning community in stochastic and adversarial online learning problems and more recently in deep learning methods. The method’s full-matrix incarnation offers much better theoretical guarantees and potentially better empirical performance than its diagonal version; however, this version is computationally prohibitive and so the simpler diagonal version often is used in practice. We introduce a new method, CompAdaGrad, that navigates the space between these two schemes and show that this method can yield results much better than diagonal AdaGrad while avoiding the (effectively intractable) $O(n^3)$ computational complexity of full-matrix AdaGrad for dimension n . CompAdaGrad essentially performs full-matrix regularization in a low-dimensional subspace while performing diagonal regularization in the complementary subspace. We derive CompAdaGrad’s updates for composite mirror descent in case of the squared ℓ_2 norm and the ℓ_1 norm, demonstrate that its complexity per iteration is linear in the dimension, and establish guarantees for the method independent of the choice of composite regularizer. Finally, we show preliminary results on several datasets.

1 Introduction

Modern machine learning applications often involve high-dimensional datasets with large sample sizes, on which simple algorithms such as variants of online gradient descent are competitive with more complicated batch algorithms. In addition to often being more practical computationally as compared to batch methods, online methods apply to a wider range of scenarios such as online prediction against individual sequences. In the online non-stochastic (adversarial) setting, the regret of a learning algorithm is a more natural quantity to analyze than the cumulative loss (for the latter an adversary may as well emit a data sequence of pure noise). Although vanilla online gradient descent (Zinkevich, 2003) obtains provably optimal regret in a minimax sense (Abernethy et al., 2008), if a method can simultaneously admit better guarantees against easy data sequences (such as low-dimensional data or i.i.d. stochastic data) while maintaining the fallback guarantee of minimax optimality against the nastiest data sequences, such a method is even better. Such methods fall within the recently-sculpted space of “learning faster from easy data” (Grünwald et al., 2013).

AdaGrad (Duchi et al., 2011) embodies an adaptive gradient family of algorithms, the two most prominent cases being full-matrix AdaGrad and diagonal AdaGrad. Whereas the former is computationally intractable in high-dimensions, the latter is tractable with only linear complexity in the dimension of the data. Both methods admit sequence-dependent regret bounds that can be much better than the regret bounds of online gradient descent. However, the computational levity of the diagonal version is matched with a price: if the components of the gradient are highly correlated, the diagonal version may fail to adapt well while the full-matrix version continues to be adaptive.

*Research conducted while at Australian National University

†Research conducted while at NICTA & Australian National University

In addition to its theoretical strengths, even the diagonal version of AdaGrad has exhibited strong empirical performance on real-world problems (Duchi et al., 2011). Moreover, diagonal AdaGrad also has been incorporated into deep learning algorithms that currently achieve state-of-the-art learning performance on a number of difficult image classification tasks (Dean et al., 2012).

Even though it is theoretically superior in terms of the regret, the full-matrix version has not seen similar large-scale applications due to its high per-round complexity of $O(n^3)$. This raises the question: is it possible to design a method lying between the full-matrix and diagonal versions of AdaGrad whose per-round complexity is $O(n)$? In this work, we answer this question in the affirmative, up to log factors, by presenting CompAdaGrad. This method replaces full-matrix AdaGrad’s Bregman divergence regularization with the sum of a *compressed* Bregman divergence operating in a low-dimensional subspace and a diagonal Bregman divergence operating in the orthogonal complement of that subspace. The precise form can be seen in (5). This method admits a theoretical guarantee that appears to be between full-matrix AdaGrad and diagonal AdaGrad, as shown in Section 3.

It is natural to ask why one would compress at the regularization level rather than simply compressing the data itself via a random projection from the very beginning. There are several reasons why compression at the regularization level makes more sense.

1. An initial compression of the data can reduce computational complexity but may lose information irretrievably, whereas by shifting the compression into the regularization component of the objective some part of the otherwise lost complementary information can still be exploited.
2. In an adversarial setting, if the learning algorithm commits to a fixed low-dimensional subspace, even an oblivious adversary can ensure that all the interesting action occurs in the complement of this subspace.
3. In some applications, one actually needs a predictor in the original space. This might be for interpretability, for instance.

In the next section, we review AdaGrad and derive CompAdaGrad. In Section 3 we present a regret bound for CompAdaGrad. We show how to compute updates for certain composite regularizers in Section 4. This section also contains a result that may be of independent interest: Theorem 4 establishes $O(n)$ complexity for computing the n -dimensional Walsh-Hadamard Transform of a 1-sparse vector. In Section 5, we present experimental results on several datasets. Finally, we conclude the paper.

2 AdaGrad: The Full, the Diagonal, and the Compressed

Notation. Throughout this paper, we let $\mathcal{X} = \mathbb{R}^n$ and leave extensions to convex subsets for future work. For a strongly convex, differentiable function ψ , let B_ψ be the Bregman divergence induced from ψ , defined as

$$B_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle.$$

For a sequence of subdifferentiable convex functions f_1, \dots, f_T , let g_1, \dots, g_T be a corresponding sequence of subgradients, so that for each $t \in [T]$ we have $g_t \in \partial f_t$.

Online Convex Optimization game. We consider the following online learning protocol parameterized by a convex regularization function $\varphi : \mathcal{X} \rightarrow \mathbb{R}_+$.

Let Nature be an oblivious adversary; that is, before the game begins Nature selects its sequence of functions possibly with knowledge of Learner’s (potentially randomized) strategy. The game then proceeds over a sequence of rounds:

For round $t = 1, 2, \dots$

- (1) Learner makes a prediction x_t in action space \mathcal{X} .
- (2) Nature reveals a convex loss function $f_t : \mathcal{X} \rightarrow \mathbb{R}_+$.

(3) Learner suffers composite loss $f_t(x_t) + \varphi(x_t)$.

The goal is to find a strategy that minimizes the regret, defined as follows. Let f_1, \dots, f_T be a sequence of functions chosen by an oblivious adversary. Then the regret on this sequence is

$$\mathcal{R}(f_1, \dots, f_T) = \sum_{t=1}^T (f_t(x_t) + \varphi(x_t)) - \inf_{x^* \in \mathcal{X}} \sum_{t=1}^T (f_t(x^*) + \varphi(x^*)). \quad (1)$$

AdaGrad with composite mirror descent embodies one family of learning strategies for obtaining low regret. Although previously AdaGrad also has been presented with regularized dual averaging, in this work we restrict to composite mirror descent for simplicity. AdaGrad for composite mirror descent is described by the updates

$$x_{t+1} = \arg \min_{x \in \mathcal{X}} \{ \eta \langle g_t, x \rangle + \eta \varphi(x) + B_{\psi_t}(x, x_t) \}, \quad (2)$$

for some constant learning rate $\eta > 0$ and some adaptive choice of convex function ψ_t .

We define $G_t := \sum_{s=1}^t g_s g_s^T$ and use the notation $\|x\|_A^2 = x^T A x$. In its full-matrix incarnation, AdaGrad uses the choice (Duchi et al., 2011)

$$\psi_t(x) = \frac{1}{2} \|x\|_{G_t^{1/2}}^2.$$

As shown by Duchi et al. (2011) and reproduced in Section 3 for convenience, full-matrix AdaGrad admits a strongly adaptive regret bound. Unfortunately, the update (2) for full-matrix AdaGrad is not tractable for large n because it involves a matrix square root and solving an n -dimensional linear system, each of which costs time $O(n^3)$. In response to this issue, there is a diagonal version of AdaGrad that admits updates in time $O(n)$. For a square matrix A , let $\text{diag}(A)$ be the diagonal matrix satisfying $\text{diag}(A)_{ii} = A_{ii}$ for all i and $\text{diag}(A)_{ij} = 0$ for all (i, j) such that $i \neq j$. The diagonal version of AdaGrad uses the choice (Duchi et al., 2011)

$$\psi_t(x) = \frac{1}{2} \|x\|_{\text{diag}(G_t)^{1/2}}^2.$$

This method also admits theoretical guarantees (cf. (Duchi et al., 2011) or Section 3). However, by design the diagonal method ignores the correlations between the components of the gradients.

CompAdaGrad. In this work, we introduce *CompAdaGrad*, a method which combines the full-matrix approach in a subspace plus the diagonal approach in the complementary subspace.

The idea of CompAdaGrad starts by restricting full-matrix AdaGrad’s Bregman divergence to a low dimensional subspace by way of a mapping $\Pi : \mathbb{R}^n \rightarrow \mathbb{R}^k$ for some $k \leq n$ (and typically $k \ll n$). Since $\Pi G_t \Pi^T = \sum_{s=1}^t \Pi g_s (\Pi g_s)^T$, this leads to the modified Bregman divergence term

$$\frac{1}{2} \|\Pi(x - x_t)\|_{(\Pi G_t \Pi^T)^{1/2}}^2 = \frac{1}{2} \|x - x_t\|_{\Pi^T (\Pi G_t \Pi)^{1/2} \Pi}^2, \quad (3)$$

which was also used by Krummenacher and McWilliams (2014).¹ Our first remark is that if $k = n$ and Π is in the orthogonal group, then it is easy to see that setting ψ_t as in the RHS of (3) recovers full-matrix AdaGrad.

A deficiency of (3) is that it ignores all of the action in the orthogonal complement of the image of Π ; however, this action can be addressed naturally by directly considering the action in this complementary subspace. To this end, define P to be the orthogonal projector corresponding to Π , defined as $P := \Pi^T (\Pi \Pi^T)^{-1} \Pi$. The corresponding complementary orthogonal projector is then $P^\perp := I - P$. Incorporating complementary regularization into the Bregman divergence with a diagonal approximation (since the dimension of the complementary subspace $n - k$ is presumably high) yields

$$\frac{1}{2} \|\Pi(x - x_t)\|_{(\Pi G_t \Pi^T)^{1/2}}^2 + \frac{\tau}{2} \|P^\perp(x - x_t)\|_{\text{diag}(G_t^\perp)^{1/2}}^2, \quad (4)$$

¹Those authors select Π to be an SRHT, described below, and we do this as well.

where $G_t^\perp := \sum_{s=1}^t (P^\perp g_s)(P^\perp g_s)^T = P^\perp G_T P^\perp$ and $\tau \geq 0$ is a parameter.

Note that when $k = n$ and Π is in the orthogonal group, (4) still recovers full-matrix AdaGrad (since the complementary subspace is empty). Additionally, when $k = 0$ and $\tau = 1$, (4) recovers diagonal AdaGrad since $P^\perp = I$. Modulo a small modification to ensure strong convexity of ψ_t , we have just derived the method that we call CompAdaGrad, which is characterized by setting ψ_t as

$$\psi_t(x) = \frac{1}{2} \|x\|_{A_t^{(r)} + \tau A_t^{(c)}}^2 \quad (5)$$

$$\text{for } A_t^{(r)} := \Pi^T (\Pi G_t \Pi^T + \delta_r I)^{1/2} \Pi \quad \text{and} \quad A_t^{(c)} := P^\perp \left(\text{diag}(G_t^\perp)^{1/2} + \delta_c I \right) P^\perp.$$

SRHT specialization. Throughout the rest of this paper, we take Π to be a Subsampled Randomized Hadamard Transform (SRHT), defined as $\Pi := \sqrt{\frac{n}{k}} R H \Sigma$ for a row selector $R \in \mathbb{R}^{k \times n}$, an n -dimensional (orthogonal) Walsh-Hadamard matrix H , and a diagonal Rademacher matrix $\Sigma \in \mathbb{R}^{n \times n}$ (whose diagonal entries are drawn i.i.d. as $+1$ and -1 with equal probability). The row selector R is induced from the distribution of all cardinality- k subsets of n indices. The idea behind (5) is that if there is interesting action in a low-dimensional subspace, the SRHT (i.e. the $A^{(r)}$ -part of the regularization) potentially can capture this action (Tropp, 2011). However, as the SRHT can only capture action in a low-dimensional subspace, it is critical to also include *complementary* regularization by way of the $A^{(c)}$ part of the regularization. This is especially true in the adversarial regime, where an adversary can pick up on Learner's fixed SRHT and make all of the interesting action happen in the complement. In this sense, CompAdaGrad can have the benefit of full-matrix AdaGrad under an oblivious, easy adversary who restricts most of the action in a low-dimensional subspace, while also maintaining guarantees against a harder adversary who uses their knowledge of $\text{Im}(P)$.

The choices of k and τ offer useful degrees of freedom for massive datasets. As k increases toward n , the method more closely resembles full-matrix AdaGrad, and the regret bound generally becomes stronger as k increases, while the computational complexity for various methods also becomes larger with larger k . The parameter τ on the other hand allows one to modulate how much emphasis to place on the complementary subspace, with this action increasingly ignored as τ decreases to 0. The next two sections help provide an understanding of the trade-off between regret and computation respectively.

3 Regret Bounds

Diagonal AdaGrad inherently is unable to adapt to correlations in a data sequence, as can be seen by Duchi et al.'s regret bound for this method (cf. Theorem 5 of Duchi et al. (2011)):

Theorem 1.

$$\mathcal{R}(f_1, \dots, f_T) \leq \frac{1}{2\eta} \max_{t \in [T]} \|x^* - x_t\|_\infty^2 \sum_{j=1}^n \left(\sum_{t=1}^T g_{t,j}^2 \right)^{1/2} + \eta \sum_{j=1}^n \left(\sum_{t=1}^T g_{t,j}^2 \right)^{1/2} \quad (6)$$

Although diagonal AdaGrad can perform well on high-dimensional sparse data, the above summations over the n dimensions are symptomatic of its inability to adapt to highly correlated dimensions.

In contrast, full-matrix AdaGrad admits the regret bound (cf. Theorem 7 of Duchi et al. (2011))

Theorem 2.

$$\mathcal{R}(f_1, \dots, f_T) \leq \frac{\delta}{\eta} \|x^*\|_2^2 + \frac{1}{2\eta} \max_{t \in [T]} \|x^* - x_t\|_2^2 \text{tr}(G_T^{1/2}) + \eta \text{tr}(G_T^{1/2}). \quad (7)$$

The above result, depending primarily on the trace of the covariance of the gradients, can exploit high correlations among the dimensions of points in a data sequence. We will show that, in certain situations of interest, CompAdaGrad can come close to the above regret guarantee at a small fraction of the computational complexity.

By design, it is quite straightforward to work out a regret bound for CompAdaGrad by leveraging the existing analysis of both full-matrix and diagonal AdaGrad from [Duchi et al. \(2011\)](#). Although in the context of concentration inequalities it is important to use the SRHT scaled as $\Pi = \sqrt{\frac{n}{k}} RH\Sigma$, in our analysis we instead analyze a variant of CompAdaGrad that is defined by replacing Π with the unscaled $\tilde{\Pi} := RH\Sigma$ (which notably satisfies $\tilde{\Pi}\tilde{\Pi}^T = I$). The following regret bound is for this variant of CompAdaGrad. Also, we set $\delta_r = \delta_c = \delta$ because it simplifies the presentation.

Theorem 3. *CompAdaGrad with learning rate $\eta > 0$ and $\delta > 0$ satisfies*

$$\begin{aligned} \mathcal{R}(f_1, \dots, f_T) &\leq \frac{\delta}{2\eta} \|x^* - x_1\|_2^2 \\ &\quad + \frac{1}{2\eta} \left(\max_{t \in [T]} \|P(x^* - x_t)\|_2^2 \operatorname{tr} \left((\tilde{\Pi} G_T \tilde{\Pi}^T)^{1/2} \right) + \max_{t \in [T]} \|P^\perp(x^* - x_t)\|_\infty^2 \|Z^\perp\|_{2,1} \right) \\ &\quad + \eta \left(\operatorname{tr} \left((\tilde{\Pi} G_T \tilde{\Pi}^T)^{1/2} \right) + \|Z^\perp\|_{2,1} \right), \end{aligned} \quad (8)$$

where $Z^\perp = (P^\perp g_1 \dots P^\perp g_T)^T$ and $A \mapsto \|A\|_{2,1}$ is the sum of the ℓ_2 norms of the columns of A .

The proof can be found in Appendix A.1.

Note that in the case where the data occupies a low-dimensional subspace and the SRHT preserves the action within this subspace, the above bound is similar to the regret bound for full-matrix AdaGrad (7). However, even when the data sequence is not so easy, [Theorem 3](#) still offers a fall-back guarantee based on the action in the orthogonal complement. In the event that the dimensions of the gradients are uncorrelated and the gradients do not occupy a low-dimensional subspace, the guarantee for diagonal AdaGrad provided by [Theorem 1](#) could be better than the guarantee for CompAdaGrad provided by [Theorem 3](#).

4 Computations

In this section, we show two important composite regularizers for which it is possible to compute the updates steps for CompAdaGrad: the squared ℓ_2 regularizer and the ℓ_1 regularizer. Some of the results below rely upon the following conjecture which we hope to affirm in the long version.

Conjecture 1. *The n -dimensional Walsh-Hadamard Transform of an r -sparse vector can be computed in time $O(n \log r)$.*

For any results that rely on the conjecture for some variable (e.g. r) indicating the sparsity level, we present the results instead with a *primed* version of the variable (e.g. r') with the understanding that $r' = r$ if the conjecture is true and $r' = n$ otherwise.

Regardless of the veracity of [Conjecture 1](#), we do however prove the following weaker result.

Theorem 4. *The n -dimensional Walsh-Hadamard Transform of a 1-sparse vector can be computed in time $O(n)$.*

The proof is constructive, and the algorithm and its analysis can be found in Appendix A.4.

4.1 CompAdaGrad with the squared ℓ_2 composite regularizer

An update for squared ℓ_2 composite mirror descent with the compressed Bregman divergence can be written in the form

$$\min_x \eta \langle g_t, x \rangle + \frac{1}{2} \|\Pi(x - x_t)\|_{K_t}^2 + \frac{\tau}{2} \|P^\perp(x - x_t)\|_{D_t}^2 + \frac{\eta\lambda}{2} \|x\|_2^2,$$

where $K_t = (\Pi G_t \Pi^T + \delta_r I)^{1/2}$ and $D_t = \operatorname{diag}(G_t^\perp)^{1/2} + \delta_c I$.

To simplify, we define $g := \eta g_t$, $K := K_t$, and $D = \tau D_t$, and we replace $\eta\lambda$; the above is then

$$\min_x \langle g, x \rangle + \frac{1}{2} \|\Pi(x - x_t)\|_K^2 + \frac{1}{2} \|P^\perp(x - x_t)\|_D^2 + \frac{\lambda}{2} \|x\|_2^2. \quad (9)$$

Since P is the orthogonal projector corresponding to the low-dimensional mapping Π , the above can be written equivalently as

$$\min_x \left\{ \langle Pg, x \rangle + \frac{1}{2} \|P(x - x_t)\|_{\Pi^T K \Pi}^2 + \frac{\lambda}{2} \|Px\|_2^2 + \langle P^\perp g, x \rangle + \frac{1}{2} \|P^\perp(x - x_t)\|_D^2 + \frac{\lambda}{2} \|P^\perp x\|_2^2 \right\}.$$

From the above final rewriting, it is clear that the optimization decouples over the two subspaces, with the minimizer of the above problem x^* being equal to the sum of

$$x_{\parallel}^* := \arg \min_{x \in \text{Im}(P)} \left\{ \langle Pg, x \rangle + \frac{1}{2} \|x - Px_t\|_{\Pi^T K \Pi}^2 + \frac{\lambda}{2} \|x\|_2^2 \right\} \quad (10)$$

$$\text{and} \quad x_{\perp}^* := \arg \min_{x \in \text{Im}(P^\perp)} \left\{ \langle P^\perp g, x \rangle + \frac{1}{2} \|x - P^\perp x_t\|_D^2 + \frac{\lambda}{2} \|x\|_2^2 \right\}. \quad (11)$$

As shown in Appendix A.2, the solution to (10) is

$$x_{\parallel}^* = (\Pi^T K \Pi + \lambda I)^{-1} (\Pi^T K \Pi x_t - Pg). \quad (12)$$

This can be computed efficiently as follows; the proof can be found in Appendix A.2.

Lemma 5. x_{\parallel}^* can be computed as

$$x_{\parallel}^* = \Pi^T \left(\frac{n}{k} K + \lambda I \right)^{-1} \left(K \Pi x_t - \frac{k}{n} \Pi g \right) \quad (13)$$

in time $O(n \log k' + k^3)$.

To solve (11), we take the dual. First, observe that (11) can be rewritten as

$$\begin{aligned} \min_x \quad & \langle P^\perp g, x \rangle + \frac{1}{2} \|x - P^\perp x_t\|_D^2 + \frac{\lambda}{2} \|x\|_2^2 \\ \text{s.t.} \quad & \Pi x = \mathbf{0}. \end{aligned}$$

The Lagrangian is

$$\mathcal{L}(x, \nu) = \langle P^\perp g, x \rangle + \frac{1}{2} \|x - P^\perp x_t\|_D^2 + \frac{\lambda}{2} \|x\|_2^2 + \langle \nu, \Pi x \rangle.$$

Lemma 6. Problem (11) can be solved as

$$x_{\perp}^* = (D + \lambda I)^{-1} (-P^\perp g + DP^\perp x_t - \Pi^T \nu^*) \quad (14)$$

with

$$\nu^* = (\Pi B \Pi^T)^{-1} \Pi B (-P^\perp g + DP^\perp x_t - DB y - \lambda B y + y), \quad (15)$$

for $B = (D + \lambda I)^{-1}$ and $y = -P^\perp g + DP^\perp x_t$. These computations can be completed in time $O(nk \log k)$.

The proof can be found in Appendix A.2.

4.2 CompAdaGrad with the ℓ_1 composite regularizer

We first sketch the LARS algorithm (Efron et al., 2004). We then show how the computations can be done efficiently for CompAdaGrad.

Re-expression as a LASSO problem. Using the same simplifications as in (9), the update step is

$$\arg \min_x \langle g, x \rangle + \frac{1}{2} \|x - x_t\|_A^2 + \lambda \|x\|_1.$$

for the symmetric matrix $A = \Pi^T K \Pi + P^\perp D P^\perp$. The above can be re-expressed as

$$x_{t+1} = \arg \min_x \left\{ \langle u, x \rangle + \frac{1}{2} \langle x, Ax \rangle + \lambda \|x\|_1 \right\}, \quad (16)$$

for $u = g - Ax_t$. Note that A is positive definite whenever $\delta_r, \delta_c > 0$.

The problem (16) can be expressed as a LASSO problem, as the optimal objective value is equal to

$$\begin{aligned} \min_x \frac{1}{2} \|x\|_A^2 - \langle -A^{-1}u, x \rangle_A + \frac{1}{2} \|-A^{-1}u\|_A^2 + \lambda \|x\|_1 &\equiv \min_x \frac{1}{2} \|x - (-A^{-1}u)\|_A^2 + \lambda \|x\|_1 \\ &\equiv \min_x \frac{1}{2} \|A^{1/2}x - (-A^{-1/2}u)\|^2 + \lambda \|x\|_1. \end{aligned}$$

Efficient computations The standard LARS algorithm takes as input a matrix of covariates X and targets y . In our case, $X = A^{1/2}$ is too expensive to compute since $A \in \mathbb{R}^{n \times n}$, while $y = A^{-1/2}u$ is similarly too expensive to compute. Therefore, we cannot just run the usual LARS algorithm but instead have to find efficient ways to perform some of the algorithm's internal computations.

The two computations for LARS that need to be handled differently are the computation of (a) the correlation of the covariates with the current residual; and (b) the entries of the Gram matrix for doing Cholesky Insert operations.

For both of the above, a certain k -by- k matrix Q arises in the computations which can be pre-computed and re-used for the entire run of LARS. Q is defined as $\frac{n}{k}K + RH\Sigma D\Sigma H^T$, or equivalently as $\frac{n}{k}K + RHDHR^T$. The second term can be computed in time $O(nk \log k)$, in parallel over columns, precisely like $\Pi B \Pi^T$ from (15), as explained in the proof of Lemma 6.

Given a current hypothesis β , the correlation of the covariates with the residual can be computed as $X^T(y - X\beta) = A^{1/2}A^{-1/2}u - A^{1/2}A^{1/2}\beta = u - A\beta$. Thus, a matrix vector multiplication is the main work. This operation can be completed in time $O(n \log k' + k^2)$, as shown in Appendix A.3.

When performing a Cholesky Insert (increasing the active set Λ by one), we need to compute part of a column of the Gram matrix. If the new dimension is i , we need to compute G_{ii} and G_{ij} for each $j \in \Lambda$. Since the Gram matrix $G = A$, we just need to compute $e_i^T A e_j$ for each $j \in \Lambda$.

To compute G_{ii} , observe that

$$G_{ii} = A_{ii} = d_i \langle RHe_i, QRHe_i \rangle - 2d_i \frac{k}{n}, \quad (17)$$

To compute (17), we first compute RHe_i in $O(k \log n)$ and store this for use in future rounds. The additional cost for computing $QRHe_i$ is then $O(k^2)$, and we re-use this result in (18). So the cost is $O(k \log n + k^2)$. To compute $G_{i,\Lambda}$, we compute for each $j \in \Lambda$

$$G_{ij} = \sigma_i \sigma_j (\langle RHe_j, QRHe_i \rangle - (d_i + d_j) \langle RHe_j, RHe_i \rangle). \quad (18)$$

Computing (18) can be done in $O(|\Lambda|k)$ since we have already stored RHe_j for all $j \in \Lambda$ and already computed $QRHe_i$.

5 Experiments

For all the experiments we used the squared ℓ_2 regularizer. For CompAdaGrad, we always set δ_r and δ_c to be equal and hereafter refer to them both as δ . When reporting test error for MNIST and Reuters RCV1, we select the values of δ , η , λ , and τ that attained the lowest online zero-one loss.

Method	Zero-one test risk
Diagonal	0.056084
CompAdaGrad-512	0.056155

Table 1: Reuters RCV1 test error.

Method	Zero-one test risk
Diagonal	0.2696
CompAdaGrad-256	0.2696

Table 2: Foxes and Wolves test error

MNIST 4 vs 9. This dataset was constructed by selecting 400 random prototypes (200 from each class) from the training set and using as features a Gaussian kernel computation of each data point with each of the 400 prototypes. We train on the entire training set for MNIST 4 and 9 and test on the test set for MNIST 4 and 9. The results in Fig. 1 indicate that CompAdaGrad well-outperforms the diagonal method, and as the SRHT dimension increases from 25 to 256, the gains over the diagonal method are all the larger. It is worth mentioning that the CompAdaGrad improves over diagonal AdaGrad even for $k = 25$. We suspect that that CompAdaGrad surpasses diagonal AdaGrad on this task because the features are dense and highly correlated.

Reuters RCV1. We used Reuters RCV1-v2 (Lewis et al., 2004). After stopping, stemming, and retaining only those unigrams and bigrams that occurred at least twice in the corpus, the data consisted of 800,000 documents in 1,889,478 dimensions. We computed gradients in mini-batches of 160 points and performed a hypothesis update from each resulting averaged gradient. We generated 4 random permutations of the data. For each permutation, we train on the first 75% and test on the last 25%. Each of the 4 runs is an experiment. We report the average test error over the 4 experiments. This test error need not correspond to a single $(\eta, \lambda, \delta, (\tau))$ configuration. As shown in Table 1, CompAdaGrad with $k = 512$ does not improve upon the diagonal version. We suspect the lack of improvement is because the data is very sparse and does not admit a low-dimensional linear subspace capturing a large amount of the action in the data. Therefore, it does not help much to pick up on correlations.

Foxes and Wolves. We grabbed a visual-bag-of-words representation for 829 images of kit foxes (*Vulpes macrotis*) and 1,156 images of red wolves (*Canis rufus*) from ImageNet (Russakovsky et al., 2015) (originally from Flickr). The original features are 1000 visual words which represent 1000 clusters of a random subset of 10 million SIFT features. We computed visual word bigrams by selecting² a radius of 0.002 for visual words to be considered collocated (based on (x, y) coordinates for the visual words); this led to 238,822 bigrams which occurred at least once in the 1,985 images. The final dataset for learning consisted of the unigram counts and bigram counts, constituting 239,822 features.

We trained on the first 75% of a random permutation of the data and tested on the remainder. Because the number of instances in this dataset is small relative to the dimension, the online zero-one loss is not a sensible rule for parameter tuning as a relatively large number hypotheses contributing to the online loss are from early rounds where not much learning has taken place. We therefore tuned the parameters based on the online training zero-one loss of the final hypothesis of the online learning algorithm. The results presented in Table 2 are inconclusive. Both diagonal AdaGrad and CompAdaGrad with $k = 256$ obtain the same zero-one test risk. This either could be due to the hardness of the task, the small sample size, or the fact that the features are still quite sparse due to the visual bag-of-words bigram representation.

6 Conclusion

We have introduced a new adaptive gradient method, CompAdaGrad, which can obtain regret bounds competitive with full-matrix AdaGrad under easier adversaries that put most of the interesting action in a low-dimensional subspace, while maintaining a fallback regret bound in the case that the adversary is not easy. CompAdaGrad also admits efficient updates for certain choices of the composite regularizer. In preliminary experiments, we demonstrate that CompAdaGrad can sometimes obtain lower risk as compared

²We selected 0.002 based on the quantiles of the distance distribution of all pairs of visual words in images.

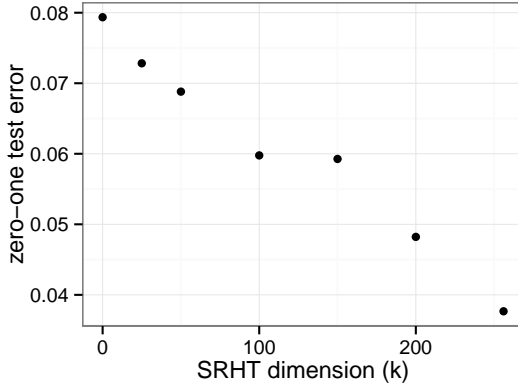


Figure 1: MNIST 4 vs 9 test error.

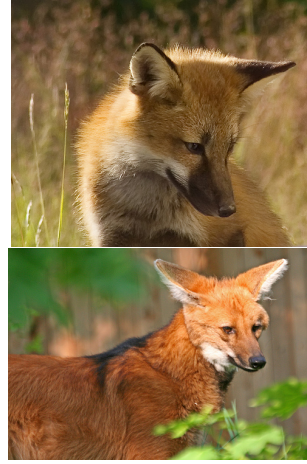


Figure 2: A kit fox and a red wolf

to diagonal AdaGrad, but this outcome appears to be tied to the density and more importantly the correlation of the features. In the future, we intend to apply CompAdaGrad to massive datasets with dense, highly correlated features. Along these lines, one promising application for future work is to incorporate CompAdaGrad into the training of deep learning methods.

Acknowledgments

This work is supported in part by the Australian Research Council Discovery Project DP0987773 and NICTA, which is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

References

- Jacob Abernethy, Peter L Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the nineteenth annual conference on computational learning theory*, 2008.
- Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, et al. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*, pages 1223–1231, 2012.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.
- Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
- Peter Grünwald, Wouter M. Koolen, and Alexander Rakhlin, editors. *NIPS Workshop on "Learning faster from easy data"*, 2013.
- Gabriel Krummenacher and Brian McWilliams. RadaGrad: Random projections for adaptive stochastic optimization. In *OPT 2014: 7th NIPS Workshop on Optimization for Machine Learning*, 2014.
- David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015. doi: 10.1007/s11263-015-0816-y.
- Joel A Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.
- Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *AAAI*, 2003.

A Proofs

A.1 Proof of regret bound

Proof of Theorem 3. We begin similar to Proposition 3 of [Duchi et al. \(2011\)](#):

$$\begin{aligned} & \eta(f_t(x_t) + \varphi(x_{t+1}) - f_t(x^*) - \varphi(x^*)) \\ & \leq B_{\psi_t}(x^*, x_t) - B_{\psi_t}(x^*, x_{t+1}) - B_{\psi_t}(x_{t+1}, x_t) + \eta\langle \eta^{-1/2}(x_t - x_{t+1}), \eta^{1/2}g_t \rangle. \end{aligned}$$

In our case, ψ_t is defined as

$$\psi_t(x) := \frac{1}{2}\|x\|_{\tilde{\Pi}^T((\tilde{\Pi}G_t\tilde{\Pi}^T)^{1/2}+\delta I)\tilde{\Pi}}^2 + \frac{1}{2}\|x\|_{P^\perp(D_t^{1/2}+\delta I)P^\perp}^2, \quad (19)$$

with D_t is the diagonal matrix defined by the diagonal entries of $P^\perp G_t P^\perp$.

We bound the red and green parts in turn. For readability, we define $\tilde{G}_t := \tilde{\Pi}G_t\tilde{\Pi}^T$.

Red part The red part can be rewritten as

$$\begin{aligned} & -\frac{1}{2}\|\tilde{\Pi}P(x_t - x_{t+1})\|_{\tilde{G}_t^{1/2}+\delta I}^2 + \eta\langle \eta^{-1/2}P(x_t - x_{t+1}), \eta^{1/2}Pg_t \rangle \\ & -\frac{1}{2}\|P^\perp(x_t - x_{t+1})\|_{D_t^{1/2}+\delta I}^2 + \eta\langle \eta^{-1/2}P^\perp(x_t - x_{t+1}), \eta^{1/2}P^\perp g_t \rangle, \end{aligned}$$

which it is easy to verify is equal to

$$\begin{aligned} & -\frac{1}{2}\|\tilde{\Pi}P(x_t - x_{t+1})\|_{\tilde{G}_t^{1/2}+\delta I}^2 + \eta\langle \eta^{-1/2}\tilde{\Pi}P(x_t - x_{t+1}), \eta^{1/2}\tilde{\Pi}Pg_t \rangle \\ & -\frac{1}{2}\|P^\perp(x_t - x_{t+1})\|_{D_t^{1/2}+\delta I}^2 + \eta\langle \eta^{-1/2}P^\perp(x_t - x_{t+1}), \eta^{1/2}P^\perp g_t \rangle \end{aligned}$$

The above can be bounded from above by the Fenchel-Young inequality as

$$\begin{aligned} & -\frac{1}{2}\|\tilde{\Pi}P(x_t - x_{t+1})\|_{\tilde{G}_t^{1/2}+\delta I}^2 + \frac{1}{2}\|\tilde{\Pi}P(x_t - x_{t+1})\|_{\tilde{G}_t^{1/2}+\delta I}^2 + \frac{\eta^2}{2}\|\tilde{\Pi}Pg_t\|_{(\tilde{G}_t^{1/2}+\delta I)^{-1}}^2 \\ & -\frac{1}{2}\|P^\perp(x_t - x_{t+1})\|_{D_t^{1/2}+\delta I}^2 + \frac{1}{2}\|P^\perp(x_t - x_{t+1})\|_{D_t^{1/2}+\delta I}^2 + \frac{\eta^2}{2}\|P^\perp g_t\|_{(D_t^{1/2}+\delta I)^{-1}}^2, \end{aligned}$$

which is just

$$\frac{\eta^2}{2}\|\tilde{\Pi}g_t\|_{(\tilde{G}_t^{1/2}+\delta I)^{-1}}^2 + \frac{\eta^2}{2}\|P^\perp g_t\|_{(D_t^{1/2}+\delta I)^{-1}}^2.$$

Summing the bound due to the red part for $t = 1$ to T yields the following two bounds: First, from Lemma 10 of [Duchi et al. \(2011\)](#) we have

$$\sum_{t=1}^T \|\tilde{\Pi}g_t\|_{(\tilde{G}_t^{1/2}+\delta I)^{-1}}^2 \leq \sum_{t=1}^T \|\tilde{\Pi}g_t\|_{\tilde{G}_t^{-1/2}}^2 \leq 2 \operatorname{tr} \left((\tilde{\Pi}G_T\tilde{\Pi}^T)^{1/2} \right).$$

Next, from Lemma 4 of [Duchi et al. \(2011\)](#) we have

$$\sum_{t=1}^T \|P^\perp g_t\|_{(D_t^{1/2}+\delta I)^{-1}}^2 \leq \sum_{t=1}^T \|P^\perp g_t\|_{D_t^{-1/2}}^2 \leq 2 \sum_{j=1}^n \left(\sum_{t=1}^T [P^\perp g_t]_j^2 \right)^{1/2}.$$

Having sufficiently bounded the red part, we now turn to the green part.

Green part We again sum from $t = 1$ to T , yielding

$$\begin{aligned}
& \sum_{t=1}^T (B_{\psi_t}(x^*, x_t) - B_{\psi_t}(x^*, x_{t+1})) \\
&= B_{\psi_1}(x^*, x_1) + \sum_{t=1}^{T-1} B_{\psi_{t+1}}(x^*, x_{t+1}) - \sum_{t=1}^T B_{\psi_t}(x^*, x_{t+1}) \\
&\leq B_{\psi_1}(x^*, x_1) + \sum_{t=1}^{T-1} (B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1})). \tag{20}
\end{aligned}$$

Recalling the form of ψ_t in (19) which has a “ P ” part and “ P^\perp ” part, we can decompose our analysis of (20) similarly.

We first analyze the “ P ” part of (20):

$$\begin{aligned}
& \frac{1}{2} \|P(x^* - x_1)\|_{\tilde{\Pi}^T(\tilde{G}_1^{1/2} + \delta I)\tilde{\Pi}}^2 \\
&+ \frac{1}{2} \sum_{t=1}^{T-1} \left(\|P(x^* - x_{t+1})\|_{\tilde{\Pi}^T(\tilde{G}_{t+1}^{1/2} + \delta I)\tilde{\Pi}}^2 - \|P(x^* - x_{t+1})\|_{\tilde{\Pi}^T(\tilde{G}_t^{1/2} + \delta I)\tilde{\Pi}}^2 \right) \\
&= \frac{1}{2} \|P(x^* - x_1)\|_{\tilde{\Pi}^T(\tilde{G}_1^{1/2} + \delta I)\tilde{\Pi}}^2 \\
&+ \frac{1}{2} \sum_{t=1}^{T-1} \left(\|P(x^* - x_{t+1})\|_{\tilde{\Pi}^T\tilde{G}_{t+1}^{1/2}\tilde{\Pi}}^2 - \|P(x^* - x_{t+1})\|_{\tilde{\Pi}^T\tilde{G}_t^{1/2}\tilde{\Pi}}^2 \right) \\
&\leq \frac{1}{2} \|P(x^* - x_1)\|_{\tilde{\Pi}^T(\tilde{G}_1^{1/2} + \delta I)\tilde{\Pi}}^2 \\
&+ \frac{1}{2} \sum_{t=1}^{T-1} \|P(x^* - x_{t+1})\|_2^2 \lambda_{\max} \left(\tilde{\Pi}^T(\tilde{G}_{t+1}^{1/2} - \tilde{G}_t^{1/2})\tilde{\Pi} \right) \\
&\leq \frac{1}{2} \|P(x^* - x_1)\|_{\tilde{\Pi}^T(\tilde{G}_1^{1/2} + \delta I)\tilde{\Pi}}^2 \\
&+ \frac{1}{2} \sum_{t=1}^{T-1} \|P(x^* - x_{t+1})\|_2^2 \text{tr} \left(\tilde{\Pi}^T(\tilde{G}_{t+1}^{1/2} - \tilde{G}_t^{1/2})\tilde{\Pi} \right) \\
&\leq \frac{1}{2} \|P(x^* - x_1)\|_{\tilde{\Pi}^T(\tilde{G}_1^{1/2} + \delta I)\tilde{\Pi}}^2 \\
&+ \frac{1}{2} \max_{t \in [T]} \|P(x^* - x_t)\|_2^2 \sum_{t=1}^{T-1} \text{tr} \left(\tilde{\Pi}^T(\tilde{G}_{t+1}^{1/2} - \tilde{G}_t^{1/2})\tilde{\Pi} \right) \\
&\leq \frac{1}{2} \|P(x^* - x_1)\|_{\tilde{\Pi}^T(\tilde{G}_1^{1/2} + \delta I)\tilde{\Pi}}^2 \\
&+ \frac{1}{2} \max_{t \in [T]} \|P(x^* - x_t)\|_2^2 \text{tr} \left(\tilde{\Pi}^T \tilde{G}_T^{1/2} \tilde{\Pi} \right) - \frac{1}{2} \|P(x^* - x_1)\|_2^2 \text{tr} \left(\tilde{\Pi}^T \tilde{G}_1^{1/2} \tilde{\Pi} \right) \\
&\leq \frac{\delta}{2} \|P(x^* - x_1)\|_2^2 + \frac{1}{2} \max_{t \in [T]} \|P(x^* - x_t)\|_2^2 \text{tr} \left(\tilde{\Pi}^T \tilde{G}_T^{1/2} \tilde{\Pi} \right) \\
&= \frac{\delta}{2} \|P(x^* - x_1)\|_2^2 + \frac{1}{2} \max_{t \in [T]} \|P(x^* - x_t)\|_2^2 \text{tr} \left(\tilde{G}_T^{1/2} \tilde{\Pi} \tilde{\Pi}^T \right) \\
&= \frac{\delta}{2} \|P(x^* - x_1)\|_2^2 + \frac{1}{2} \max_{t \in [T]} \|P(x^* - x_t)\|_2^2 \text{tr} \left(\tilde{G}_T^{1/2} \right),
\end{aligned}$$

where the last inequality follows because $\tilde{\Pi}^T \tilde{\Pi} = P$, and the last equality follows since $\tilde{\Pi} \tilde{\Pi}^T = I$.

We now turn to the “ P^\perp ” part of (20):

$$\begin{aligned}
& \frac{1}{2} \|P^\perp(x^* - x_1)\|_{D_1^{1/2} + \delta I}^2 \\
& + \frac{1}{2} \sum_{t=1}^{T-1} \left(\|P^\perp(x^* - x_{t+1})\|_{D_{t+1}^{1/2} + \delta I}^2 - \|P^\perp(x^* - x_{t+1})\|_{D_t^{1/2} + \delta I}^2 \right) \\
& = \frac{1}{2} \|P^\perp(x^* - x_1)\|_{D_1^{1/2} + \delta I}^2 \\
& + \frac{1}{2} \sum_{t=1}^{T-1} \left\langle P^\perp(x^* - x_{t+1}), (D_{t+1}^{1/2} - D_t^{1/2}) P^\perp(x^* - x_{t+1}) \right\rangle \\
& \leq \frac{1}{2} \|P^\perp(x^* - x_1)\|_{D_1^{1/2} + \delta I}^2 \\
& + \frac{1}{2} \sum_{t=1}^{T-1} \|P^\perp(x^* - x_t)\|_\infty^2 \langle D_{t+1}^{1/2} - D_t^{1/2}, \mathbf{1} \rangle \\
& \leq \frac{1}{2} \|P^\perp(x^* - x_1)\|_{D_1^{1/2} + \delta I}^2 \\
& + \frac{1}{2} \max_{t \in [T]} \|P^\perp(x^* - x_t)\|_\infty^2 \sum_{t=1}^{T-1} \langle D_{t+1}^{1/2} - D_t^{1/2}, \mathbf{1} \rangle \\
& \leq \frac{1}{2} \|P^\perp(x^* - x_1)\|_{D_1^{1/2} + \delta I}^2 \\
& + \frac{1}{2} \max_{t \in [T]} \|P^\perp(x^* - x_t)\|_\infty^2 \langle D_T^{1/2}, \mathbf{1} \rangle - \frac{1}{2} \|P^\perp(x^* - x_1)\|_\infty^2 \langle D_1^{1/2}, \mathbf{1} \rangle \\
& \leq \frac{\delta}{2} \|P^\perp(x^* - x_1)\|_2^2 + \frac{1}{2} \max_{t \in [T]} \|P^\perp(x^* - x_t)\|_\infty^2 \langle D_T^{1/2}, \mathbf{1} \rangle.
\end{aligned}$$

The final regret bound Putting everything above together yields the final bound

$$\begin{aligned}
& \sum_{t=1}^T (f_t(x_t) + \varphi(x_{t+1}) - f_t(x^*) - \varphi(x^*)) \\
& \leq \frac{\delta}{2\eta} \|x^* - x_1\|_2^2 \\
& + \frac{1}{2\eta} \left(\max_{t \in [T]} \|P(x^* - x_t)\|_2^2 \operatorname{tr} \left((\tilde{\Pi} G_T \tilde{\Pi})^{1/2} \right) + \max_{t \in [T]} \|P^\perp(x^* - x_t)\|_\infty^2 \langle D_T^{1/2}, \mathbf{1} \rangle \right) \\
& + \eta \left(\operatorname{tr} \left(\tilde{G}_T^{1/2} \right) + \langle D_T^{1/2}, \mathbf{1} \rangle \right)
\end{aligned}$$

with

$$\langle D_T^{1/2}, \mathbf{1} \rangle = \sum_{j=1}^n \left(\sum_{t=1}^T [P^\perp g_t]_j^2 \right)^{1/2}.$$

□

A.2 Proof of computational results

First, we establish that (12) is in fact the solution to the constrained problem (10).

We rely on the following claim:

$$(\Pi^T K \Pi + \lambda I)^{-1} \Pi^T = \Pi^T (K \Pi \Pi^T + \lambda I)^{-1}. \quad (21)$$

To see this, multiply both sides on the left by $(\Pi^T K \Pi + \lambda I)$, yielding

$$\begin{aligned}\Pi^T &= (\Pi^T K \Pi + \lambda I) \Pi^T (K \Pi \Pi^T + \lambda I)^{-1} \\ &= \Pi^T (K \Pi \Pi^T + \lambda I) (K \Pi \Pi^T + \lambda I)^{-1} = \Pi^T.\end{aligned}$$

Now, observe that (12) is the solution to the unconstrained version of (10). Thus, it is sufficient to show that $P^\perp (\Pi^T K \Pi + \lambda I)^{-1} (\Pi^T K \Pi x_t - P g)$ is equal to zero. This is indeed true since

$$\begin{aligned}P^\perp (\Pi^T K \Pi + \lambda I)^{-1} (\Pi^T K \Pi x_t - P g) &= \left(I - \frac{k}{n} \Pi^T \Pi \right) (\Pi^T K \Pi + \lambda I)^{-1} \left(\Pi^T K \Pi x_t - \frac{k}{n} \Pi^T \Pi g \right) \\ &= \left(I - \frac{k}{n} \Pi^T \Pi \right) (\Pi^T K \Pi + \lambda I)^{-1} \Pi^T \left(K \Pi x_t - \frac{k}{n} \Pi g \right) \\ &= \left(I - \frac{k}{n} \Pi^T \Pi \right) \Pi^T (K \Pi \Pi^T + \lambda I)^{-1} \left(K \Pi x_t - \frac{k}{n} \Pi g \right) \\ &= (\Pi^T - \Pi^T) (K \Pi \Pi^T + \lambda I)^{-1} \left(K \Pi x_t - \frac{k}{n} \Pi g \right) \\ &= 0.\end{aligned}$$

Proof of Lemma 5. First,

$$(\Pi^T K \Pi + \lambda I)^{-1} (\Pi^T K \Pi x_t - P g) = (\Pi^T K \Pi + \lambda I)^{-1} \Pi^T \left(K \Pi x_t - \frac{k}{n} \Pi g \right) \quad (22)$$

since $P = \Pi^T (\Pi \Pi^T)^{-1} \Pi = \frac{k}{n} \Pi^T \Pi$.

Using (21), we have that the RHS of (22) is equal to

$$\Pi^T (K \Pi \Pi^T + \lambda I)^{-1} \left(K \Pi x_t - \frac{k}{n} \Pi g \right) = \Pi^T \left(\frac{n}{k} K + \lambda I \right)^{-1} \left(K \Pi x_t - \frac{k}{n} \Pi g \right). \quad \square$$

Proof of Lemma 6. We first derive the expressions for x_\perp^* and ν^* and then establish the computational complexity result.

Proof of expressions for x_\perp^* and ν^* . At the maximum the partial gradient of the Lagrangian WRT x must be zero, and so

$$\frac{\partial \mathcal{L}}{\partial x} = \mathbf{0} = P^\perp g + D(x - P^\perp x_t) + \lambda x + \Pi^T \nu.$$

Thus, we have the relation

$$x = (D + \lambda I)^{-1} (-P^\perp g + D P^\perp x_t - \Pi^T \nu) \quad (23)$$

which can be computed in time $O(n \log k)$ assuming $O(n \log k)$ computation of the Walsh-Hadamard transform of a k -sparse vector.

Define $y := -P^\perp g + D P^\perp x_t$ and $B := (D + \lambda I)^{-1}$. Plugging in this expression for x into the Lagrangian yields the dual problem

$$\max_\nu \langle P^\perp g, B(y - \Pi^T \nu) \rangle + \frac{1}{2} \|B(y - \Pi^T \nu) - P^\perp x_t\|_D^2 + \frac{\lambda}{2} \|B(y - \Pi^T \nu)\|_2^2 + \langle \nu, \Pi B(y - \Pi^T \nu) \rangle.$$

which is equivalent to the problem

$$\max_\nu -\langle P^\perp g, B \Pi^T \nu \rangle + \frac{1}{2} \|B(y - \Pi^T \nu) - P^\perp x_t\|_D^2 + \frac{\lambda}{2} \|B(y - \Pi^T \nu)\|_2^2 + \langle \nu, \Pi B y \rangle - \|\nu\|_{\Pi B \Pi^T}^2. \quad (24)$$

Let us verify that the objective is concave. The Hessian is

$$\Pi B D B \Pi^T + \lambda \Pi B^2 \Pi^T - 2 \Pi B \Pi^T = \Pi (B D B + \lambda B^2 - 2B) \Pi^T.$$

The term on the RHS sandwiched between Π and Π^T expands to

$$(D + \lambda I)^{-2} D + \lambda (D + \lambda I)^{-2} - 2(D + \lambda I)^{-1},$$

which (by way of the diagonal structure) easily works out to be $-(D + \lambda I)^{-1}$. Hence, the problem is strongly concave.

Thus, we have reduced the problem to a low-dimensional concave unconstrained problem which can be solved analytically. Differentiating *just the first-order (in ν) terms* of the objective of (24) WRT ν yields

$$\begin{aligned} & \Pi B P^\perp g + \Pi B D B y - \Pi B D P^\perp x_t + \lambda \Pi B^2 y - \Pi B y \\ &= \Pi B (P^\perp g + D B y - D P^\perp x_t + \lambda B y - y). \end{aligned}$$

Finally, setting the gradient to zero and solving for ν yields

$$\nu = (\Pi (D + \lambda I)^{-1} \Pi^T)^{-1} \Pi B (-P^\perp g + D P^\perp x_t - D B y - \lambda B y + y). \quad (25)$$

We can then compute x using (23).

Proof of computational complexity. First, the computation for (15) can be done efficiently because the vector to which the inverse is applied can be computed in $O(n \log k')$. To see this, observe that for $x \in \mathbb{R}^n$, we have $Px = \Sigma H R^T R H \Sigma x$, which involves (in sequence) scaling by Σ , a trimmed Walsh-Hadamard Transform (WHT) in $O(n \log k)$, applying R^T to create a k -sparse n -dimensional vector in $O(n)$, application of a WHT to this k -sparse vector in $O(n \log k')$, and a final scaling by Σ .

Computing the linear system matrix (involved in the inverse) is the most expensive step: the columns of this k -by- k matrix can be computed in parallel, and each column can be computed in time $O(n \log k)$ since, for $j \in [k]$ we have, for some $i \in [n]$, that $\Pi B \Pi^T = R H (B (H e_i))$. Now, this computation involves (in sequence) a WHT applied to a 1-sparse vector in time $O(n)$ as per Theorem 4, a scaling by B , and a trimmed WHT in $O(n \log k)$. Hence, the entire matrix can be computed in time $O(nk \log k)$.

Finally, the linear system can be solved in $O(k^3)$, yielding a total complexity of $O(nk \log k + k^3)$. \square

A.3 Matrix-vector multiplication

We do the computation in the order indicated in the final line below.

$$\begin{aligned} A\beta &= (\Pi^T K \Pi + P^\perp D P^\perp) \beta \\ &= \left(\frac{n}{k} \Sigma H R^T K R H \Sigma + (I - \Sigma H R^T R H \Sigma) D (I - \Sigma H R^T R H \Sigma) \right) \beta \\ &= D\beta - D \Sigma H R^T R H \Sigma \beta + \Sigma H R^T (Q R H \Sigma \beta - R H \Sigma D \beta) \\ &= D\beta - D \Sigma H (R^T (\tilde{\Pi} \beta)) + \Sigma H (R^T (Q (\tilde{\Pi} \beta) - \tilde{\Pi} (D \beta))). \end{aligned} \quad (26)$$

For a vector x , $\tilde{\Pi}x$ can be computed in $O(n \log k)$ using the SRHT. Diagonal scaling (by D and Σ) costs $O(n)$. Applying Q to some $z \in \mathbb{R}^k$ costs $O(k^2)$. Finally, for $z \in \mathbb{R}^k$, observe that $H R^T z$ is the Walsh-Hadamard transform of a k -sparse vector (since $R^T z$ scatters the k entries of z into a k -sparse n -dimensional vector), which can be computed in time $O(n \log k')$.

A.4 Pseudo-code and Complexity of product of Walsh-Hadamard matrix with one-sparse vector

The following algorithms calculates $r = H v$, where $H \in \mathbb{R}^{n \times n}$ is a Walsh-Hadamard matrix, v a 1-sparse vector, and $n = 2^m$ because all Walsh-Hadamard matrices have dimensions which are powers of

2. Assume the indices of v and H are zero based and that i is the index of the only non-zero component of v . The algorithm utilizes the fact that the result is equal to the i -th column of H , denoted h , multiplied by the scalar $v(i)$.

The column h of H can be found by using the recursive structure of the Walsh-Hadamard matrix

$$H_n = \begin{bmatrix} H_{n/2} & H_{n/2} \\ H_{n/2} & -H_{n/2} \end{bmatrix}. \quad (27)$$

If the index i is in the interval $0 \leq i < \frac{n}{2}$ then the column h will sit in the left part of the block matrix in (27) and therefore the sub-vector of the first $\frac{n}{2}$ components of h will be equal to the sub-vector of the second $\frac{n}{2}$ components, $h(\frac{n}{2}, \dots, n-1) = h(0, \dots, \frac{n}{2}-1)$. If however i is in the other interval $\frac{n}{2} \leq i < n$ then the column h will be in the right part of the block matrix of (27) and therefore the second sub-vector is the negative of the first sub-vector, $h(\frac{n}{2}, \dots, n-1) = -h(0, \dots, \frac{n}{2}-1)$. Testing which of the two cases occurs can be efficiently done by checking the value of bit m of index i , where we use the convention that the least-significant bit of i is bit 0.

The problem has now been reduced to calculating the components $h(0, \dots, \frac{n}{2}-1)$ in the matrix $H_{n/2}$ which can be done similarly by looking at the value of bit $m-1$ of index i . The base case is reached when we set $h(0) = 1$ and then update according to

$$h(1) = \begin{cases} +h(0) & \text{if bit 0 of } i \text{ is } 0 \\ -h(0) & \text{if bit 0 of } i \text{ is } 1 \end{cases}. \quad (28)$$

The result r can therefore be constructed by initializing the first component of r to $v(i)$ and then copying $1 + 2 + \dots + \frac{n}{2} = n-1$ components of sub-vectors of the result r into yet uninitialized sub-vectors of r using the bits of the index i to choose the appropriate sign. The complexity is therefore $O(n)$.

Algorithm 1: Product of Walsh-Hadamard matrix with 1-sparse vector

Input:	n $i \in \{0, \dots, n-1\}$ $v(i)$	the dimension of the vector v index of single non-zero component of v value of non-zero component of v
---------------	--	--

Output: $r(0, \dots, n-1)$ where $r = Hv$

$r(0) \leftarrow v(i)$

$m \leftarrow 1$

while $m < n$ **do**

if $i \bmod 2 = 0$ then	$r(m, \dots, 2*m-1) \leftarrow +r(0, \dots, m-1)$
else	$r(m, \dots, 2*m-1) \leftarrow -r(0, \dots, m-1)$
end if	
$i \leftarrow i/2$	
$m \leftarrow m*2$	

end while
